Thanks for many comments ! Below, reviewers shown by their CMT # (①–④) (excerpt of comments in *blue*), reference to lines of submitted draft as `Lxxx`, (OP) = Open Problem(s), (SM) = Supplementary Material, (CR) = Camera Ready. ▷ ①②③. Our title originates in Vapnik's quote: `https://tinyurl.com/y5jnurau`. Roughly, Vapnik's Structural Risk Minimization (as in the pictured equation) embeds uncertainty due to both sampling and model choice – justifying Vapnik's quote. **However** the loss in SRM is ad-hoc and as we write, statistical decision theory has long made an *intensional* case for its choice. Our approach suggests that a Bayesian framework *could* be better than the frequentist ones [NM20] to cope with this uncertainty – hence our title. We take the title feedback seriously and can elaborate in the additional CR page OR opt for a technical title, even when the connection with Vapnik would wane in this change.

▷ ① *[...] benefit from extra space [...]*: Thanks, we plan to use the extra CR page as detailed in this rebuttal. *[...] distribution of ν [...]*: This is the approximate posterior over source functions and is conditional on our model in the usual Bayesian manner; the r.h.s. merely features the parameters of the posterior mean function – *c.f.* Lemma 7.

▷ ② *While both the research [...] proper losses can meet*: Our method does work with **any** proper loss in which the source is embedded. Brier score, log-loss (*our experimental choice*). ②'s impossibility concern highlights a key feature and is covered by our Theorem 3. We gratefully offer rebuttal-size argument restricted to ②'s examples, which are symmetric proper (SP) with invertible links. 2 steps: (i) elicit composite link χ such that $(-\underline{L}_{\mathrm{us}})^\star(-y\chi(u)) = (-\underline{L}_{\textcircled{2}})^\star(-yu), \forall y \in \{-1,1\}, u \in [0,1]$. Invertible link implies $(-\underline{L})^\star$ strictly monotonic and thus invertible, and we get $\chi(u) = -y\cdot((-\underline{L}_{\mathrm{us}})^\star)^{-1}\circ(-\underline{L}_{\textcircled{2}})^\star(-yu) = -((-\underline{L}_{\mathrm{us}})^\star)^{-1}\circ(-\underline{L}_{\textcircled{2}})^\star(-u), \forall y \in \{-1,1\}$. The last identity holds because SP losses satisfy $(-\underline{L})^\star(-x) = (-\underline{L})^\star(x) - x$ [NN08, eq (10)]. (ii) we get the source ν from χ and $\underline{L}_{\mathrm{us}}$ using eq. (3) in our paper, which can be expressed by an universal kernel. QED. *Also [...] only work on a binary setting*: We **respectfully disagree**: it works without modification in multiclass multilabel case by using a 1-vs-all or 1-vs-1 multiple coding. *Although this paper [...] check ceratin metrics like expectation calibration [...] proposed approach*: There is probably a **misunderstanding** here. Consistency, calibration, rates are **formal** properties of a loss. For example, calibration is equivalent to a negative derivative in zero of the margin loss, which guarantees "label consistency". It is not an experimental property. We are happy to make L96-L97 and ref. [BJM06] more explicit using CR. Only robustness could be checked but it would fairly deserve a paper of its own; we are happy to push it as OP. *[...] it seems to be a complex and expensive method [...]*: We **respectfully disagree**: *c.f.* Sec 5 of [NM20] – their *fastest* option is $\mathcal{O}(N \log N)$ *per iter* but at the expense of a very involved data structure ($\mathcal{O}(N^2)$ without). Our simpler to implement $\mathcal{O}(M^2N)$ per EM iter (L529) admits control of M, a meaningful knob to drive complexity below [NM20].

▷ ③ *[...] a clear algorithmic breakdown [...]*: The paper being already dense in content, we are happy to push additional pseudo-code in SM. *I wonder [...] interesting insights*: We integrate out ν for prediction similarly to a GP classifier. *[...] M Aitkin [...] debate [...]*: Thanks ! See the motive behind our title above as a token that inference on the loss + model could save a few "Bayesian eggs" (we can discuss in CR). *I wonder about the need to deal with [the ISGP] [...]*: We give extensive references to alternatives on L52-L55, all of which involve certain drawbacks compared to our ISGP. However, we believe that the reviewer's suggestion is both novel and very interesting for this application, thanks ! *[...] worth stating [...] parametric approximation of a GP*: Agreed, thanks (although much GP inference exploits a parametric approximation) . *I find it unclear why Sec 4.1[...]*: Agreed on the poor headings. Here univariate ≈ ISGP & multivariate ≈ our complete model; we will improve for CR. *The experiments do not seem [...] One idea [...]*: We respectfully disagree wrt the "learning the loss" problem: see Table 1 in [NM20] and justification in our L34-L35. **However** we do agree that building up above GLMs is probably the best way to further widen the gap, *yet* one must keep in mind the complexity cost, so it is more than about more complicated models (see ▷ ② above). *[...] inference method [...] outdated* : Agreed, however efficient variational inference for the ISGP is an OP due to intractable integrals.

▷ ④ (ref tokens much appreciated) *A1*: This is an interesting OP, thanks; our flexiblility in the loss may ease the burden of underlying model capacity. *W2*: The purpose of Th.3 is to show that kernels **do** fit to solve our problem. Considering the complexity constraint for all approaches (see ▷ ② above), one might benefit to yield on universality for practical purposes – and our experiments display that this is more than fine, in line with our intuition given the vast richness of the space of losses we consider. *W1+W3*: The universality of the trigonometric kernel is an OP, **but** we do offer an alternative Nyström based method in Appendix F which we show to be readily applicable to the (universal) Gaussian kernel (satisfying the conditions of Theorem 3). Moreover, pending the integral of L583, it could be not just the Gaussian kernel that would be covered. Plus, we may have slightly undersold the Nyström method from a complexity standpoint (at least form a theoretical standpoint): given the univariate-ness of the ISGP, one length scale and one output scale parameter should suffice as the kernel hyper-parameters. Then, even a finite difference approximation of the marginal likelihood derivatives should be within an (albeit rather large) constant factor of the corresponding computation time for the trigonometric kernel. *W4*: Agreed, great pointer ! **But** we caution that those eigen-functions depend on the hyper-parameters, making inference more expensive as above. *C1, C2, C3* Agreed, thanks ! *C4*: Well spotted for the GP, thanks ! **But** of course inapplicable to our ISGP. *C5*: See comments on experiments and W6 above – we will use the additional page to alleviate dependencies on SM, and we shall add refs in Sec 6.